Metagenomic evidence for coexistence of SARS and H1N1 in patients from 2007-2012 flu seasons

Qi Liu[1,2,4]†, Zhenglin Du[2,3,4]†, Sihui Zhu[2,3,4], Wenming Zhao[2,3,4], Hua Chen[1,2,4]*, Yongbiao Xue[2,3,4]*

[1] CAS Key Laboratory for Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China.

[2] China National Center for Bioinformation, Beijing 100101, China.

[3] National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China.

[4] University of Chinese Academy of Sciences, Beijing 100049, China.

†These authors contributed equally.

*Corresponding author. Email: ybxue@big.ac.cn; chenh@big.ac.cn

Abstract

By re-analzying public metagenomic data from 101 patients infected with influenza A virus during the 2007-2012 H1N1 flu seasons in France, we identified 22 samples with SARS-CoV sequences. In 3 of them, the SARS genome sequences could be fully assembled out of each. These sequences are highly similar (99.99% and 99.7%) to the artificially constructed recombinant SARS-CoV (SARSr-CoV) strains generated by the J. Craig Venter Institute in USA. Moreover, samples from different flu seasons have different SARS-CoV strains, and the divergence between these strains cannot be explained by natural evolution. Our study also shows that retrospective studies using public metagenomic data from past major epidemic outbreaks serves as a genomic strategy for the research of origins or spread of infectious diseases.

Genome sequencing has been used to identify pathogen, trace virus origin, and provide outbreak surveillance for infectious disease studies. The availability of complete genome of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in January 2020 sped up the identification of the pathogen and facilitated the development of effective vaccines. With the accumulation of extensive metagenomic sequence data in the past years, one potential application is to carry out genome-based retrospective studies of major historical outbreaks to understand the occurrence and development of viral epidemics.

Here we re-analysed public metagenomic data from 101 patients infected with influenza A virus (H1N1), collected by the Pasteur Institute in France between 2007 and 2012, spanning five consecutive flu seasons [1]. In 22 (21.78%) out of the 101 patient samples, we identified genomic fragments of SARS-associated coronavirus (SARS-CoV) with different proportions (0.0003% ~ 0.6127%) (Table S1). For 3 samples with sequencing depth >30 and genome coverage >99% (ERR1091908, ERR1091910, and ERR1091914), we were able to assemble the complete SARS genomes (Table S1). These genomes have different numbers of mutations ranging from 10 to 116 compared with the SARS-CoV reference genome (NC_004718.3, see Table S1). In addition, the samples collected during the same flu season share similar mutations, while samples from different flu seasons possess different sets of mutations (Figure S1, Table 1), suggesting that distinct SARS-CoV strains existed in different flu seasons in France.

SARS-CoV once caused an outbreak of SARS in 2002, a life-threatening respiratory infectious disease [2], but disappeared in human populations after 2003. To investigate the origin of these SARS-CoV sequences detected in the patients infected with influenza A virus between 2007 and 2012, the three assembled SARS-CoV genome sequences (ERR1091908, ERR1091910 and ERR1091914) were pooled together with 250 complete SARS-CoV genomes downloaded from NCBI database. We constructed a phylogeny of these sequences with the neighbor-joining approach and a haplotype network with the median-joining approach, respectively (Figure 1). We found that ERR1091908 and ERR1091910 are clustered with the SARS-CoV ExoN1 strain (colored in blue in Figure 1) while ERR1091914 is clustered with the SARS-CoV wtic-MB strain (colored in black in Figure 1), consistent with the finding of differential SARS-CoV mutations in the patient samples from different flu seasons. Note that SARS-CoV ExoN1, SARS-CoV wtic-MB, SARS-CoV MA15 and SARS-CoV MA15 ExoN1 all belong to recombinant SARSr-CoV, a group of SARS-CoV sequences artificially constructed using the same infectious clone (ic) recombinant virus strain of SARS-CoV Urbani (AY278741) that was originally isolated from a patient with SARS-CoV [3-5]. Moreover, the three newly assembled SARS-CoV sequences along all the recombinant SARSr-CoV sequences are separated from other naturally occurring SARS-CoV sequences including AY278741 (colored in yellow and cyan in Figure 1) in the phylogeny and haplotype network. Taken together, these results indicate that all the three SARS-CoV sequences (ERR1091908, ERR1091910 and ERR1091914) are more likely to be categorized as artificially constructed recombinant SARSr-CoV strains.

ERR1091914 is almost identical (99.99%) to SARS-CoV wtic-MB (19 identical sequences in the data), with only one base pair different (bp) (R10626A, referred to the genome position of FJ882938.1) after trimming the 20-bp at the 5'-end and 44-bp at the 3'-end of ERR1091914 (Table 1). Here, R represents A or G base. There are a total of 15 mutation differences between ERR1091914 and AY278741, the closest naturally occurring SARS-CoV, and the 15 mutations are shared by all the recombinant SARSr-CoV, indicating that ERR1091914 is indeed derived from the SARSr-CoV wtic-MB sequences instead of evolving independently from naturally occurring SARS-CoV sequences.

The other two newly identified SARS-CoV sequences (ERR1091908 and ERR1091910) are almost identical to each other with only one base different (99.99%). Both sequences are highly similar (99.70%) to the SARS-CoV ExoN1 sequence FJ882941.1, with 88 base differences after trimming 21-bp at the 5'-end and 65-bp at 3'-end (Table 1). Annotation results showed that 84 of the 88 bases locate at coding regions of SARS-CoV and result in 45 amino acid (AA) substitutions. Most of these mutations are in the coding regions of open reading frame 1a (ORF1a) (ORF1a) (13 substitutions), ORF1b (7 substitutions) and the spike (S) glycoprotein (9 substitutions, Table 2), among which ORF1a and S glycoprotein are functionally related to the increase of viral virulence, transmission and pathogenicity [6], suggesting potential gain-function effects of these mutations in ERR1091908 and ERR1091910. SARS-CoV ExoN1 strains are known to have a 21-fold increase in mutation rate during replication in previous research [5], which is consistent with the fact that the sequences within the SARS-CoV ExoN1 clades of the haplotype network are highly divergent compared with other recombinant SARSr-CoV clades (Figure S2). Assuming a mutation rate of 1.0 × 10^{-3} per site per year for naturally evolving SARS-CoV and the SARS-CoV genome length of 29,751 bp, only 1.69 months (50.7 days) are needed to generate 88 site differences between the two SARS-CoV ExoN1 sequences, indicating a recent divergence time between the two sequences (ERR1091908 and ERR1091910) and FJ882941.1.

The sequencing data of 101 patients were submitted by the Pasteur Institute in France, which established a laboratory with level-three biosafety standard and conducted research on SARS-CoV [7-9]. All the 116 recombinant SARSr-CoV sequences were submitted by the J. Craig Venter Institute (JCVI) from Tennessee, USA. The two institutes have collaborated and published their work on viral genome sequence [10].

Discussion

One possible explanation of the co-existence of SARS and H1N1 sequences in the patients is that the artificially constructed recombinant SARSr-CoV caused a co-infection outside the laboratory during 2007 - 2012, but did not result in SARS-CoV epidemic; an alternative hypothesis is a contamination of the samples in the lab since the Pasteur Institute also conducted SARS-CoV studies. Moreover, samples from different flu seasons have different strains of SARS-CoV, and the divergence between these SARS-CoV strains cannot be explained by natural evolution. This intriguing finding warrants further efforts to sleuth out the culprit. In 2014 it was reported that the Pasteur Institute France once lost vials containing patient samples collected during SARS (https://www.sciencemag.org/news/2014/05/frances-institut-pasteur-under-fire-over-missing-sars-vials). It raises a serious concern of laboratory biosafety in both institutions. Our study also shows that retrospective studies using public metagenomic data from past major epidemic outbreaks serve as a useful genomic strategy for the research of origins or spread of infectious diseases.

Materials and methods

Data collection

The metagenomic data of 101 patients infected by the influenza A virus were downloaded from NCBI SRA database (project ID: PRJEB11406). These samples were collected by National influenza center (France) near Paris between 2007 and 2012, spanning 5 consecutive flu seasons. Sequencing data of these samples were submitted by Institute Pasteur, France in 2018 (Table S1).

Variant calling and consensus sequence generating

After removing sequencing adapters and trimming consecutive low-quality bases from both the 5' and 3' read ends using cutadapt [11], clean reads were mapped to the SARS-CoV-2 genome (NC_045512.2) using BWA (V0.7.12) [12] with default parameters. The Picard program (http://picard.sourceforge.net) was used to sort mapping results to BAM format and mark duplicates of PCR amplification. Then GATK (V4.1.6.0) [13] was used for SNP and indel calling. Consensus sequences were generated by applying VCF variants to the reference sequence using bcftools.

Sequence alignment

250 genomes of SARS-CoV with genome length larger than 29,0000 bases, were downloaded from NCBI Nucleotide database (https://www.ncbi.nlm.nih.gov/nuccore). We then constructed a multiple sequence alignment of 253 genomes using MAFFT v7.453 [14] and the final alignment contains 30,327 nucleotides.

Phylogenetic and network analysis

Neighbor-joining (NJ) phylogenetic trees of the 253 genome sequences were constructed using MEGA 10.1.8 with default arguments [15]. Phylogenetic relationships and mutations occurred among unique genomes were further inspected from 253 genomes through median-joining networks [16] using Network 10 (http://www.fluxus-engineering.com/) to examine changes of genetic variations across places and through times. For network analysis, a 81-bp block at the 5’-end including gaps and a 77-bp block at the 3’-end including gaps and the poly-A tails in the alignment were trimmed out and the final alignment contains 30,169 nucleotides.

Pairwise sequence alignment

We used the BLAST online tools with default parameters (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to align two sequences.

Interest statement

Authors declare that they have no competing interests.

Acknowledgements

This project was supported by the National Natural Science Foundation of China (Grant No. 31571370, 91731302, and 31772435).

References

[1] Pelletier I, Rousset D, Enouf V, *et al*. Highly heterogeneous temperature sensitivity of 2009 pandemic influenza a(h1n1) viral isolates, northern france. Euro surveillance : bulletin Européen sur les maladies transmissibles, 2011, 16: 19999

[2] Ksiazek TG, Erdman D, Goldsmith CS, *et al.* A novel coronavirus associated with severe acute respiratory syndrome. New England Journal of Medicine, 2003, 348: 1953-1966

[3] Yount B, Curtis KM, Fritz EA, *et al*. Reverse genetics with a full-length infectious cdna of severe acute respiratory syndrome coronavirus. Proc Natl Acad U S A, 2003, 100: 12995-13000

[4] Roberts A, Deming D, Paddock CD, *et al.* A mouse-adapted sars-coronavirus causes disease and mortality in balb/c mice. PLoS Pathogens, 2007, 3: e5

[5] Eckerle LD, Becker MM, Halpin RA, *et al.* Infidelity of sars-cov nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. PLoS Pathogens, 2010, 6: e1000896

[6] Chinese SARS Molecular Epidemiology Consortium. Molecular evolution of the sars coronavirus during the course of the sars epidemic in china. Science, 2004, 303: 1666-1669

[7] Mantke OD, Schmitz H, ., Herve Z, *et al.* Quality assurance for the diagnostics of viral diseases to enhance the emergency preparedness in europe. Euro surveillance : bulletin Européen sur les maladies transmissibles, 2005, 10: 1-2

[8] Yap YL, Zhang XW, Andonov A, *et al.* Structural analysis of inhibition mechanisms of aurintricarboxylic acid on sars-cov polymerase and other proteins. Computational Biology & Chemistry, 2005, 29: 212-219

[9] Fontanet A. [cross-species transmission: Last obstacle before pandemic]. Transfusion Clinique Et Biologique Journal De La Société Franaise De Transfusion Sanguine, 2007, 14: 16-17

[10] Eppinger M, Rosovitz MJ, Fricke WF, *et al.* The complete genome sequence of yersinia pseudotuberculosis ip31758, the causative agent of far east scarlet-like fever. PLoS Genetics, 2007, 3: e142

[11] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet Journal, 2011, 17: 10-12

[12] Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics, 2009, 25: 1754-1760

[13] Mckenna A, Hanna M, Banks E, *et al.* The genome analysis toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. Genome Research, 2010, 20: 1297-1303

[14] Katoh K, Misawa K, Kuma K-i, *et al.* Mafft: A novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Research, 2002, 30: 3059-3066

[15] Sudhir K, Glen S, Li M, *et al.* Mega x: Molecular evolutionary genetics analysis across computing platforms. Molecular Biology & Evolution, 2018, 35: 1547-1549

[16] Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Molecular Biology & Evolution, 1999, 16: 37-48

Figure legends

Figure 1. Phylogeny tree and haplotype network of 253 SARS-CoV genome sequences. **A**. A phylogeny of the 253 SARS-CoV genome sequences is constructed with the neighbor-joining approach. **B**. Haplotype network of the 253 SARS-CoV genome sequences is constructed with the median-joining approach.

Figures

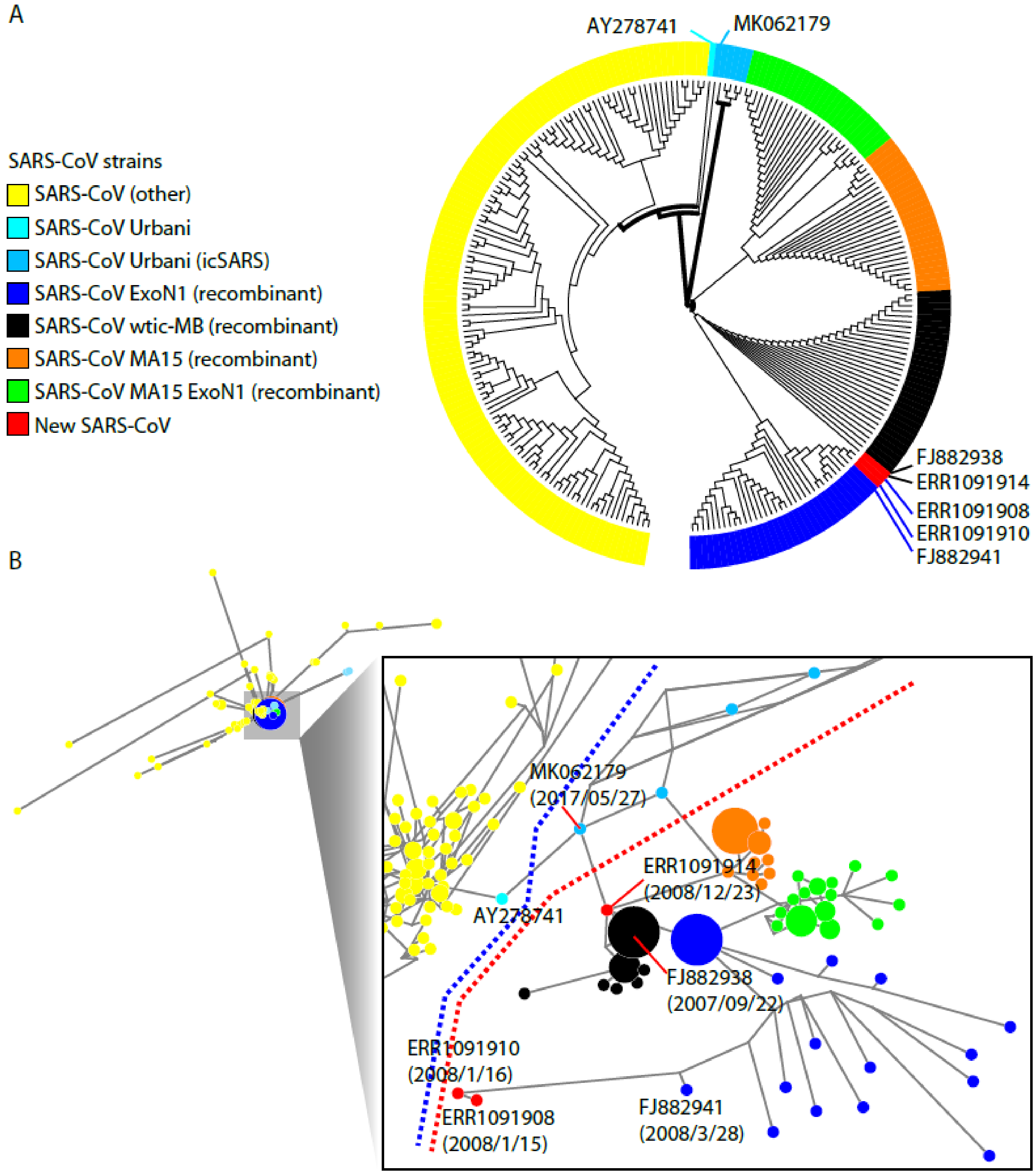

Figure 1. Phylogeny tree and haplotype network of 253 SARS-CoV genome sequences. **A**. A phylogeny of the 253 SARS-CoV genome sequences is constructed with the neighbor-joining approach. **B**. Haplotype network of the 253 SARS-CoV genome sequences is constructed with the median-joining approach.

Tables

Table 1. Summary of three SARS-CoV sequences and their closest sequences.

Cluster	Sample ID	Sample collection location	Collection date	Organism	Submitter	Identity with new SARS-CoV sequence (%)
1	ERR1091914	Haute Normandie, France	2008/12/23	Influenza A virus	Institute Pasteur	
	FJ882938	Tennessee, USA	2007/9/22	SARS-CoV wtic-MB	J. Craig Venter Institute	99.99
2	ERR1091908	Lorraine, France	2008/1/15	Influenza A virus	Institute Pasteur	
	ERR1091910	Picardie, France	2008/1/16	Influenza A virus		
	FJ882941	Nashville, Tennessee, USA	2008/3/28	SARS-CoV ExoN1	J. Craig Venter Institute	99.70

*One sequence closed to ERR1091914 is shown and the remaining 18 sequences closed to ERR1091914 are presented in Table S2.

Table 2. The positions of 84 base differences between ERR1091908 and FJ882941.1

ORF	CDS	Mutations found in ERR1091908 compared to FJ882941.1		Function
		Positions of nucleotide change (number)	Positions of amino acid change in SARS-CoV protein (number)	
ORF 1a	265-13413	654,707,1771,1905,2976,3229,3491,3603,3845,4731,4808,5015,5061,5236,5412,6087,6265,6459,6476,7484,8004,8922,10119,10658,12411,13149 (26)	148,503,989,1076,1194,1489,1515,1584,1658,2001,2071,2407,3465 (13)	associated with increased virulence, transmission, and pathogenicity during the epidemic
ORF 1b	265-21485	13874,13925,14178,14630,14876,15497,15605,15740,15821,15905,16356,16386,17269,17602,18238,18239,18244,18245,18749,18860,19082,19814,19917,20528,20555,20789,21038 (27)	987,997,1291,1402,1614,1616,2174 (7)	
S	21492-25259	21860,22206,22352,22423,23243,23374,23468,23518,23823,24249,24873,24910,24957 (13)	239,311,628,676,778,920,1128,1140,1156 (9)	associated with increased virulence, transmission, and pathogenicity during the epidemic
ORF 3a	25268-26092	25550,25626,25783,25800,26049 (5)	120,178,261 (3)	
ORF 3b	25689-26153	25783,25800,26049,26121 (4)	32,38,121,145 (4)	
E	26117-26347	26121,26226,26241,26335 (4)	2,37,42 (3)	
M	26398-27063			
ORF 6	27074-27265	27167,27248 (2)	32,59 (2)	
ORF 7a	27273-27641	27290,27639 (2)	123 (1)	
ORF 7b	27638-27772	27639,27648 (2)	1,4 (2)	
ORF 8a	27779-27898			
ORF 8b	27864-28118	27917 (1)		
nucleocapsid protein	28120-29388	28557,29271,29324 (3)	402 (1)	
ORF 9b	28130-28426			
Total number		84	45	